Approximation of the age distribution of cancer incidence using a mutational model

Alexandr N. Tetearing

July 6, 2021

Abstract

The approximation of the age distributions of cancer was carried out using a complex mutational model presented in our work [1]. Datasets from the American National Cancer Institute (SEER program) were used.

We approximated the data of age distributions of lung, stomach, colon and breast cancer in women; cancer of the lung, stomach, colon and prostate in men.

The average number of mutations (required for cancer formation) averaged over the four types of cancer is 5.25 mutations per cell in women and 5.5 mutations per cell in men. The average (over the four types of cancer) mutation rate is estimated as $4 \cdot 10^{-4}$ mutations per year per cell for women and $8 \cdot 10^{-4}$ mutations per year per cell for men.

This article is a continuation of work [1].

Contents

1. Mutational model of age distribution of cancers

In this article, we use the Dz(t) function obtained in [1] as a fitting function to approximate the datasets on the age distribution of cancer incidence:

$$Dz(t) = \frac{d\left(N_P\left(1 - e^{-Nz(t)}\right)\right)}{dt} \quad (1.1)$$

Here z is the number of cell mutations required for cancer to occur.

N_P is the number of people in the considered group (the group with the given age distribution of cancers).

The function $Nz(t)$ describes the number of cells (in the cell population of anatomical tissue) with the number of mutations z or more:

$$Nz(t) = N(t)\left(1 - e^{-at}\right)^z \quad (1.2)$$

The a parameter is the average number of mutations (key events) that occur in a cell per unit of time.

The $N(t)$ function is the growth function of the cell population (anatomical tissue). The function describes the number of cells in the cell population of the organ (tissue). The growth of the cell population occurs in accordance with the equations:

$$N(t) = \begin{cases} e^{bt} & \text{if } t \le t_1 \\ \frac{R_0}{k_0} - c_3\, e^{-\frac{k_0}{k_3}t} & \text{if } t > t_1 \end{cases} \quad (1.3)$$

Until time t_1 the population (number of cells) grows exponentially (the equation of unlimited growth), after the time t_1 the population is faced with a limitation in the food (energy) resource (the equation of limited growth).

Thus, the time t_1 is the time when the cell population has the maximum growth rate.

The constant b from the upper equation of system (1.3) is determined by the expression:

$$b = \frac{k_1 - k_0}{k_3} \quad (1.4)$$

The coefficients k_i are constant coefficients that characterize the interaction of cells with the environment.

The constant c_3 (this is the constant of integration) is found from the equation:

$$N(t_1) = \frac{R_0}{k_0} - c_3\, e^{-\frac{k_0}{k_3}t_1} \quad (1.5)$$

The R_0 constant determines the amount of resource available to the cell population per unit of time (at the stage of limited growth). The R_0/k_0 ratio determines the number of cells in the population at the final stage of population growth.

The a parameter (the number of key events per unit time) is related with the R_0 parameter . As the selected value R_0/k_0 decreases, the a parameter increases (if, for example, we consider that only stem cells are susceptible to mutations – that is, that is, only some part of the value of R_0/k_0 mutates). And vice versa – as R_0/k_0 increases, the a parameter decreases.

The time t_1 (the time when the cell population passes from the stage of unlimited growth to the stage with limited growth) is determined by the equality:

$$e^{b\,t_1} = \frac{R_0}{k_1} \qquad (1.6)$$

All these equations and coefficients are described in detail in book [2].

The $N(t)$ function is shown in Fig. 1.1.

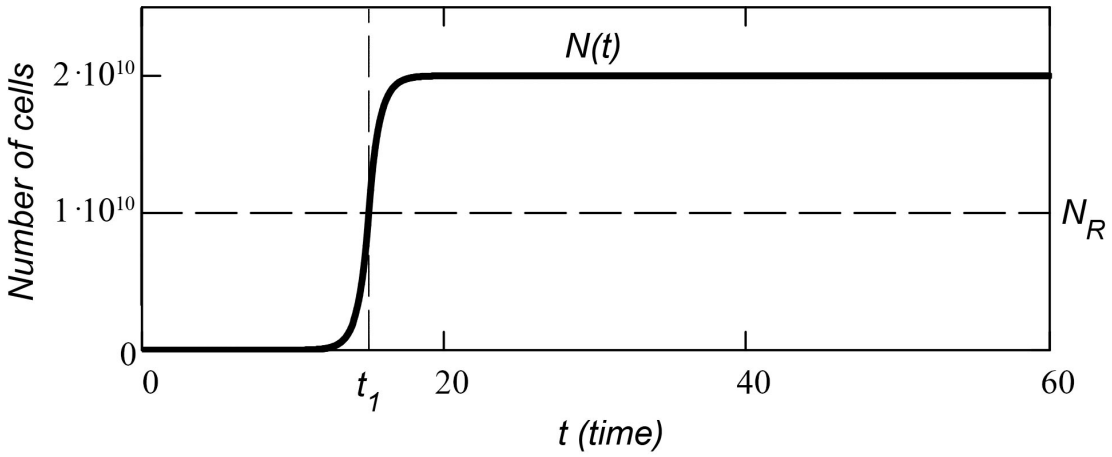

Figure 1.1. Growth function of the cell population of anatomical tissue.

For simplicity, we assume that at time $t = 0$ the cell population consists of one cell: $N(0) = 1$.

2. Datasets

We used data from the SEER (Surveillance Epidemiology and Outcomes Program) of American National Cancer as the data source [3].

These data are more detailed in comparison with the generalized data of the Belorussian register of cancers, which we used earlier in [1].

SEER data cover the age range of patients from 0 to 110-115 years. The age of initial cancer registration is recorded for all patients. Thus, we can obtain the age distribution of the incidence in one year increments.

For example, in the datasets provided by SEER, there are cases where lung cancer in women was registered at the age of 113 years, breast cancer at 118 years old; in men, there is a case when lung cancer was diagnosed in a patient at the age of 111 years.

We used data on all registered patients (separately for female and male cancers) registered from 1975 to 2016. All patients enrolled in the SEER program are considered without racial separation. The total number of cases considered is:

Women:

Colon:	362,407 cases;
Lung:	511,421 cases;
Stomach:	65,747 cases;
Breast:	1,392,873 cases;

Men:

Colon:	344,828 cases;
Lung:	654,526 cases;
Stomach:	102,000 cases;
Prostate:	1,307,624 cases.

These detailed age distributions of cancers show the shape of the distribution function very well, especially for elderly patients.

3. Approximation method

As an approximating function, we used function (1.1), which describes the mutational model of cancer, presented in [1].

As in [1], the optimization of the error function Er is carried out programmatically, by the Levenberg-Marquardt method (least squares method) for each fixed value of z, since the number of mutations z (the number of key events required for the onset of cancer in a cell) is a positive integer.

$$Er(z, a, N_P, T_S) = \frac{1}{n}\sqrt{\sum_{k=1}^{n}\left(D_k - Dz(t_k)\right)^2}$$

Here t_k is the age from the real dataset; D_k is the value of the age distribution from the real dataset for the age of t_k; $Dz(t_k)$ is the theoretical value of the age distribution for the age t_k.

The function is optimized by four parameters: z, a, N_P and T_S. The least squares method is used to select a set of parameters for which the error function $Er(z, a, N_P, T_S)$ is minimal.

We also optimize the delay parameter T_S added to formula (1.1) – we assume that cancer is diagnosed not at the moment of onset, but several years later (after the time T_S). That is, the time T_S must not be negative or equal to zero. Therefore, we substitute the difference $(t - T_S)$ into the equation for $Dz(t)$ function instead of time t.

When approximating, we used the following constant parameters: $k_0 = 0.5$; $k_1 = 1$; $t_1 = 15$. The coefficient k_3 is calculated from the equation:

$$\frac{k_3}{k_1 - k_0} ln\left(\frac{R_0}{k_1}\right) = t_1 \qquad (3.1)$$

The R_0 parameter is chosen (for each organ) so as to provide the required number of cells at the final stage of organ growth, which is determined by the anatomical size of the organ. Knowing the anatomical dimensions of the human organ and the size of the cell [4], we can estimate the number of cells $N_G = R_0/k_0$ in the cellular tissue of an adult organ.

In this work, we used the following estimates:

Colon: $N_G = 2.0 \cdot 10^8$
Lung: $N_G = 2.0 \cdot 10^{10}$
Stomach: $N_G = 2.4 \cdot 10^7$
Mammary: $N_G = 5.0 \cdot 10^7$
Prostate: $N_G = 5.6 \cdot 10^5$

To compare the error values for the age distributions of different cancers for the obtained (optimized) parameters of the Er function, we calculate the relative approximation error as a percentage of the D_{max} (maximum incidence from the dataset):

$$Er_{\%}(z, a, N_P, T_S) = \frac{100}{n} \sqrt{\sum_{k=1}^{n} \left(\frac{D_k - Dz(t_k)}{D_{max}}\right)^2} \qquad (3.2)$$

4. Results

The results are shown in Figures 4.1 and 4.2, where, for each type of cancer, the age distribution (open points in Figures 4.1 and 4.2) and theoretical age distribution curve (solid curves in Figures 4.1 and 4.2) are plotted.

The obtained parameters of the approximating function, at which the minimum approximation error is achieved, are shown in Table 4.1.

Table 4.1.
Approximation parameters for the mutational model.

Female	z	T_S	a	N_P	Er	$Er_{\%}$
Colon	6	4.25	0.00057	359,813	52.22	0.47
Lung	6	8.98	0.00030	514,798	18.44	0.10
Breast	4	6.55	0.00020	1,426,650	146.03	0.41
Stomach	5	10.41	0.00051	64,944	11.94	0.64
Averaged	**5.25**	**7.55**	**0.00040**	**591,551**	**57.16**	**0.41**

Male	z	T_S	a	N_P	Er	$Er_{\%}$
Colon	6	2.41	0.00059	346,862	24.97	0.23
Lung	5	20.56	0.00017	658,658	23.36	0.10
Prostate	5	25.54	0.00161	1,310,010	134.35	0.25
Stomach	6	0.94	0.00084	1,024,940	5.96	0.20
Averaged	**5.50**	**12.36**	**0.00080**	**835,118**	**47.16**	**0.20**

In the bottom row of Table 4.1, the parameter values averaged over four types of cancer (separately for women and men) are highlighted in bold.

The average number of mutations required to form a cancer is 5.25 mutations per cell for female cancers and 5.5 mutations per cell for male cancers.

The average number of mutations per unit time is $4 \cdot 10^{-4}$ mutations per year per cell for women and $8 \cdot 10^{-4}$ mutations per year per cell for men.

The root mean square relative error of approximation by this mutation model is 0.41 percent for women and 0.2 percent for men.

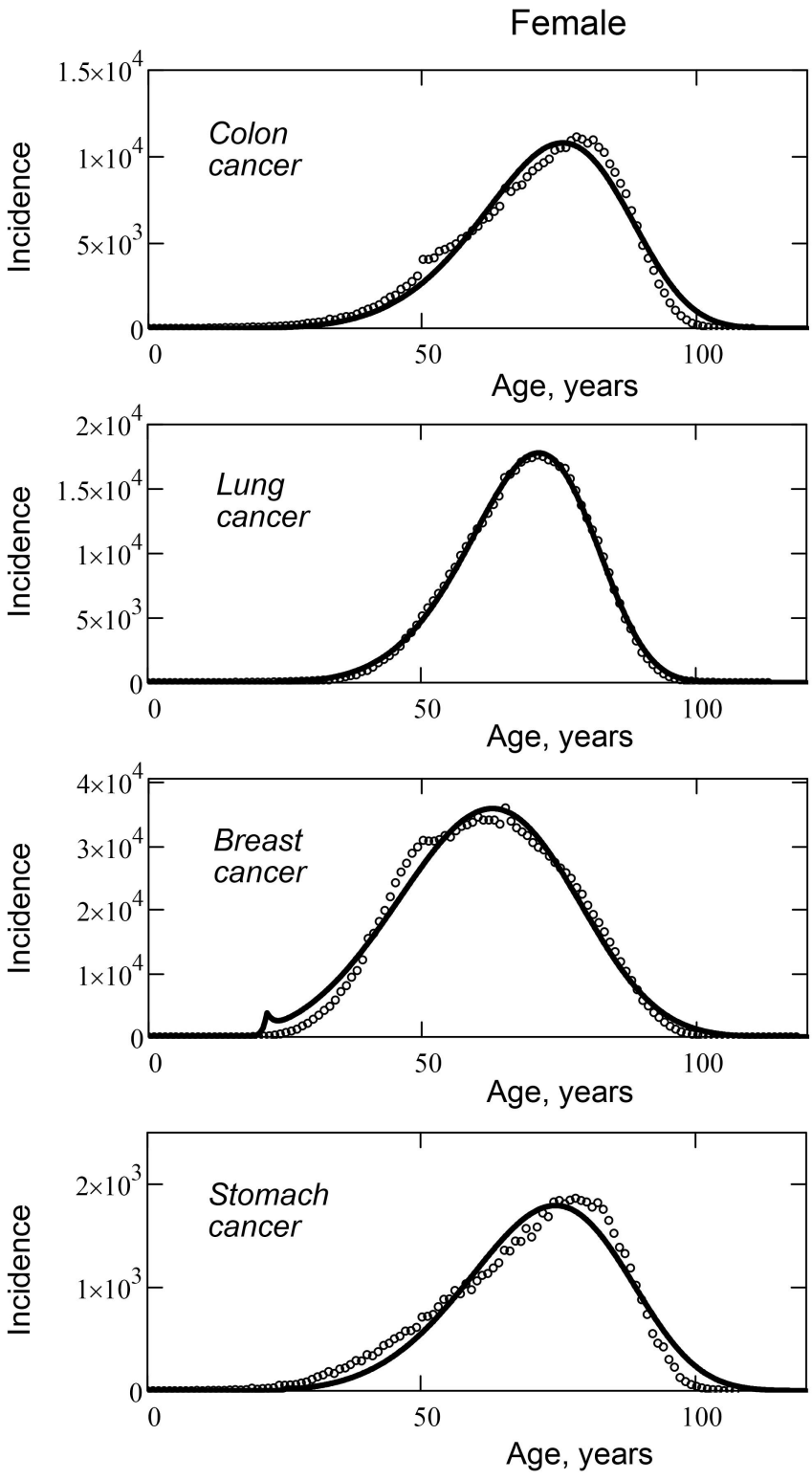

Figure 4.1. Age distribution of cancer incidence in women. The light points are real datasets; solid bold curves are approximation functions (mutational model of cancer).

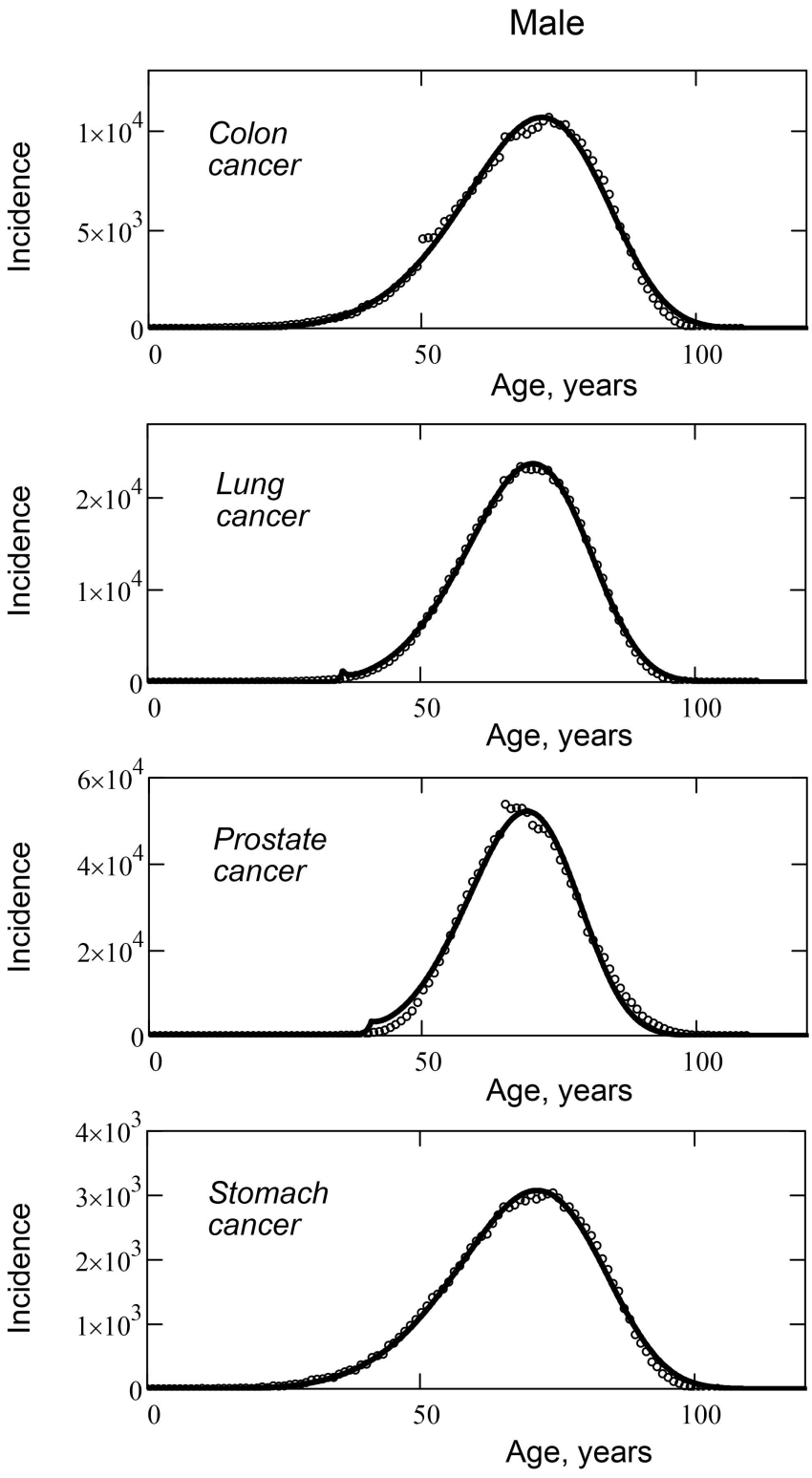

Figure 4.2. Age distribution of cancer incidence in men. The light points are real datasets; solid bold curves are approximation functions (mutational model of cancer).

5. Discussion

Figures 4.1 and 4.2 clearly show that the approximation of age distributions in men gives better accuracy than approximation of age distributions in women (light points of real data sets in men are closer to theoretical curves). The theoretical model best describes the age distribution of lung cancer (both women and men) – real data and theoretical curves are almost identical.

The average accuracy of approximation for four forms of cancer in women is two times worse than in men (bottom line, in bold, in Table 4.1, parameter $Er_{\%}$).

The mutational model used for data fitting allows us to determine the following parameters of cancer:

– the number of mutations per cell required for cancer to occur (z parameter from Table 4.1);

– the average number of mutations per unit time (year) per cell of the considered cell tissue (a parameter from Table 4.1);

– time lag in medical diagnosis of cancer (the time interval between the occurrence of cancer and its detection), (T_S parameter from Table 4.1);

Among all the presented forms of cancer (4 cancers in women and 4 cancers in men), the TS time (time of delay in diagnosis) is found only in half of the cases: these are cases of lung cancer and breast cancer in women; lung cancer and prostate cancer in men.

For other forms of cancer, with optimal parameters of the error function, the time TS turns out to be negative, and this does not correspond to reality (in the case of negative time lag, we include in table 4.1 the last non-negative value of TS, which is obtained by sequential discrete increase in the number z).

The mutational model considered in the article has significant drawbacks.

First, it should be understood that the model describes the key events that happened to a biological cell, but does not speak about the causes of these events (the model does not tell us that these events are cellular mutations).

Second, the model does not tell us why the number of such events (or mutations) is five or slightly more (bottom line of Table 4.1, in bold, z parameter)? We do not know how we can practically measure this "magic" number in any other way than by looking at the age distribution of cancers. This may indicate our lack of understanding of the processes occurring in the cell and (or) the inapplicability of this mathematical model to the processes under consideration.

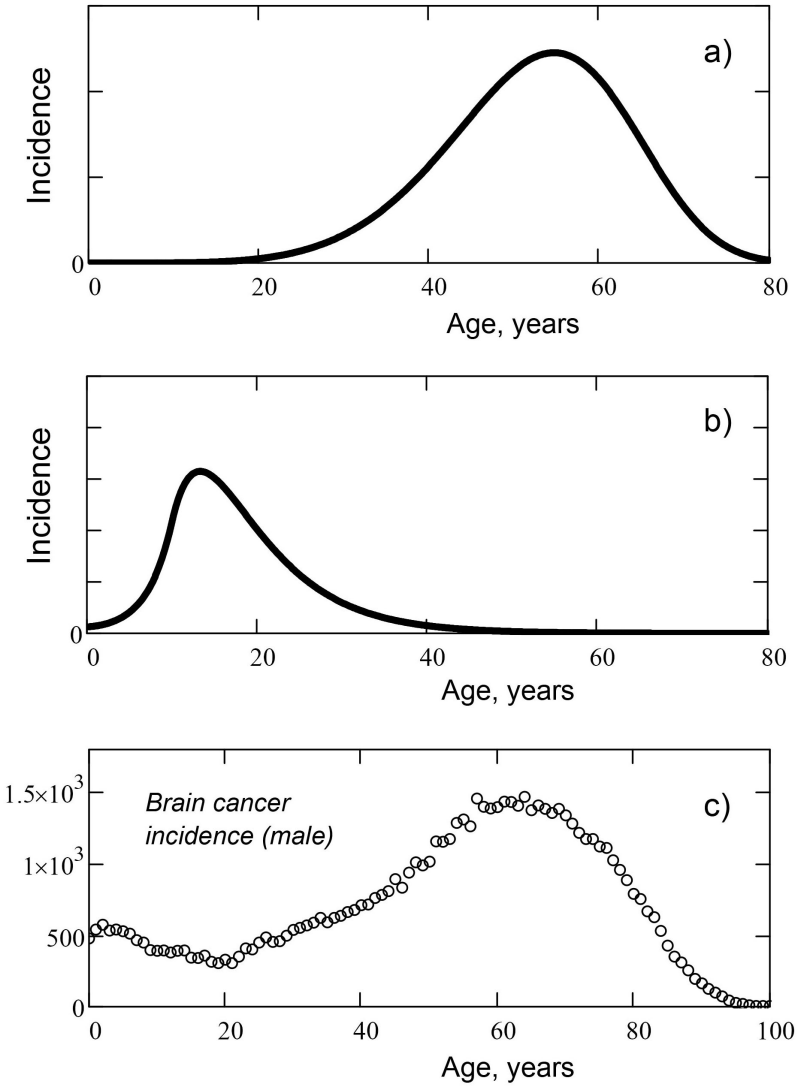

Figure 5.1. Different types of functions of age distribution of cancers: a) – mutational model of cancer; b) – model of single carcinogenic event; c) – age distribution of brain cancer in men according to the American National Cancer Institute (SEER program).

Third, the mutational model in most cases (these are cases when the number of mutations is more than two, and the time t_1 is more than 10 years) gives the age distributions that have the shape of a hill with a gentle left slope, something like that shown in Fig. 5.1-a.

But, in practice, there are other forms of age distribution of cancer. For example, Figure 5.1-c shows the age distribution of brain cancer in males (SEER data). The distribution has a local minimum to the right of the origin (between the zero and the maximum of the function).

The mutational model presented in this article cannot provide an age distribution with this shape.

Perhaps the graph in Fig. 5.1-c illustrates the fact that two carcinogenic processes occur simultaneously (but with the different intensities) in the anatomical tissue. One process gives the age distribution shown (roughly) in Fig. 5.1-a. Another process gives the age distribution shown (roughly) in Fig. 5.1-b. The sum of these distributions (for certain parameters) can give the graph shown in Fig. 5.1-c.

Therefore, other mathematical models should also be considered to describe the age distribution of cancer. It is possible that we will try to present some of these models in one of the following articles.

Saint-Petersburg
2021

References

[1] **Tetearing A.N.**
Cancer models and statistical analysis of age distribution of cancers. 2021. https://dx.doi.org/10.2139/ssrn.3841599

[2] **Tetearing A.N.**
Theory of populations. SSO Foundation, Moscow, 2012.

[3] **National Cancer Institute, Surveillance Epidemiology and End Results** Overview of the SEER Program, https://seer.cancer.gov/about/overview.html

[4] **Great Medical Encyclopedia.**
Edition 3, Moscow, 1974-1989 (in Russian).